https://doi.org/10.31852/EMF.33.2020.005.014

EVALUATING THE EFFECT OF DATA PROCESSING TECHNIQUES ON INDOOR AIR QUALITY ASSESSMENT IN BUDAPEST

Bushra Atfeh [(1)], Erzsébet Kristóf [(2)], Róbert Mészáros [(1)], Zoltán Barcza [(1,2,3)]

[(1)] Department of Meteorology, Eötvös Loránd University, H-1117 Budapest, Hungary
[(2)] Excellence Center, Faculty of Science, Eötvös Loránd University, H-2462 Martonvásár, Hungary
[(3)] Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, 165 21 Prague, Czech Republic
e-mail: bushra.at@hotmail.com

Abstract

This study focuses on indoor air quality measurements carried out in an apartment in the suburban region of Budapest. The measurements were made by an IQAir AirVisual Pro air quality monitor which is a so-called low-cost sensor capable to monitor $PM_{2.5}$ and carbon dioxide concentration. In this study we analyze data measured during January 2017 that was characterized by an extreme air pollution episode in Budapest. The aim of the study was to calculate daily indoor $PM_{2.5}$ concentrations that are comparable with the outdoor concentrations provided by the Hungarian Air Quality Monitoring Network. Given the fact that AirVisual Pro provides data with irregular sampling frequency, data processing is expected to affect the calculated daily mean concentrations. The main purpose of the study was to evaluate the effect of data processing technique selection on the calculated daily data. The results indicated that the uneven sampling frequency characteristic to AirVisual Pro indeed causes problems during data processing and has effect on the calculated means. We propose a 'best method' for data processing for sensors with irregular sampling frequency.

1. Introduction

Given the fact that we spend considerable amount of time in residential buildings and in other closed environment (workplace, school, shops, etc.), exposure to indoor air pollutants is a major issue worldwide. Indoor air quality is strongly affected by outdoor conditions (Leung, 2015; Burnett et al., 2018; WHO, 2018), and by indoor activities like cooking, use of cleaning chemicals, cosmetics, etc. (Majd et al., 2019). Despite air quality regulations and legislation, many regions are still affected by poor air quality conditions worldwide (Joss et al., 2017). In Hungary (and in Central Europe in general) residential heating is a major air pollution source as many citizens use wood and garbage for heating. Among other pollutants, Particulate Matter (PM) originating from residential heating represents a huge risk to public health that also affects indoor air quality (Kistler et al., 2012; EMEP, 2020).

In response to this challenge posed by the indoor air quality issues, low-cost sensors (LCS) monitoring $PM_{2.5}$ and CO_2 (and other gases in some cases) in indoor conditions are gaining popularity. LCSs are often criticized because of their poor accuracy and/or precision, and the lack of responsiveness to episodic high PM concentrations (Li et al., 2020; Zamora et al., 2020). Indeed, some of the sensors seem to be very problematic such as Speck LCS (Zamora et al., 2020) or the SDI sensor that provides biased data due to high and low relative humidity (Bulot et al., 2020; Tagle et al., 2020). It means that some of the LCSs are not applicable to indoor exposure estimations without post-processing. However, there are LCSs with a good overall performance which means that they might be capable to capture high pollution episodes and

the overall background PM indoor concentration and those LCSs might be promising tools in terms of future health concerns reduction (Lowther et al., 2019).

AirVisual Pro is an optical LCS that offers visualization of indoor or outdoor $PM_{2.5}$, carbon dioxide (CO_2) mixing ratio, relative humidity and temperature in real time and offers remote access from the mobile phone application. Besides, it offers option for comparison of the observed $PM_{2.5}$ with the nearest official air quality monitoring station. All these features facilitate the indoor air quality monitoring and tracking that is supported by user-friendly suggestions on the screen of the node to maintain good quality air indoors in terms of $PM_{2.5}$ and fresh air (in terms of CO_2 mixing ratio).

A few studies evaluated the quality of the AirVisual Pro observations against some reference instrument. In 2017 three units of AirVisual Pro were evaluated by Feenstra et al. (2019) for 2 months against a Met One Beta Attenuation Monitor (BAM; Met One, USA) that is an U.S. EPA designated Class III FEM (EQPM-0308-170) reference instrument. The R^2 values were around 0.7, while the mean bias was 1.3 μg m^{-3}, and the RMSE was about 6.3 μg m^{-3}. In 2018, $PM_{2.5}$ readings from AirVisual Pro were compared with a Met One 1020 Beta Attenuation Monitor (MetOne, USA), a Dust Monitor by GRIMM (model EDM180, Ainring, Germany), and a T640 PM Mass Monitor (Teledyne API, USA) reference instruments at the Air Quality Sensor Performance Evaluation Center (AQ-SPEC, 2018). The AirVisual Pro exhibited good correlations with the reference instruments at 5-minute, hourly and daily resolution as well (R^2 varied between 0.66 and 0.89) and good precision for the readings. Zamora et al. (2020) evaluated AirVisual Pro for one year in indoor conditions against a pDR (personal DataRAM™ pDR-1200, ThermoFisher Scientific, USA) reference instrument. The results showed high accuracy and high correlation with pDR (R^2 was 0.89) for a non-smoking indoor environment.

Besides calibration issues, data processing represents another source of uncertainty of the LCSs data interpretation. As some sensors are 'black boxes' (which means that the hardware and the software are not documented in detail) it is not trivial to choose the 'best method' for post-processing. As an example, AirVisual Pro's performance is affected by uncertainty related to the non-constant measurement frequency that is characteristic to the sensor.

The aim of the present study was to evaluate the effect of data processing techniques on the exposure calculations for $PM_{2.5}$ using an AirVisual Pro sensor.

2. Materials and methods

2.1. AirVisual Pro

AirVisual Pro (*Fig. 1*) – manufactured by IQAir (Switzerland) – is a low-cost sensor that offers measurements for indoor and outdoor pollution. This device provides real-time $PM_{2.5}$ (μg m^{-3} units; effective range: 0.3–2.5 μm in size), CO_2 (effective range: 400–10,000 ppm), temperature (effective range: –10 to 40 °C) and relative humidity (effective range: 0–95%) data that updates in near-real-time for continuous monitoring (IQAir website, 2020a).

AirVisual Pro is equipped with an AVPM25b optical sensor (also developed by IQAir) that measures $PM_{2.5}$ concentrations, and a SenseAir S8 (Model SE-0031, SenseAir, Sweden) sensor measuring CO_2 concentrations (Zamora et al., 2020). The $PM_{2.5}$ sensor is an advanced light-scattering laser sensor which measures the size of microscopic particulate matter (IQAir website, 2020b).

According to the manufacturer, the frequency of measurements can be set in three different modes: the custom mode which offers measurements in regular intervals (from 3 minutes to 1 hour), the continuous mode making readings in every 10 seconds, and the default mode which makes 4 readings per hour if the device is inactive. However, the default mode and the custom mode have a special feature that needs attention. If the device is actively used (it means that the

screen is activated/deactivated, or the screen content is changed) then the measurement frequency will increase and readings will be made every 10 seconds for some time before reverting back to 15-minute sampling frequency (IQAir website, 2020c). This means that the device can provide data with non-uniform frequency.

The Department of Meteorology, Eötvös Loránd University has 5 AirVisual Pro sensors. Two of them were purchased in late 2016 or beginning of 2017, which means that they can be considered as first generation AirVisual nodes. In these cases the software only provides 2 modes: the default mode and the continuous mode. In normal circumstances the continuous mode should not be used to improve the lifetime of the internal pump which means that they are operated in default mode. The other three instruments were purchased later so they belong to the new generation of the sensors in terms of software/hardware environment. It means that they can also be operated in custom mode. Nevertheless, if the AirVisual node is operated in custom mode using e.g. 3 minutes sampling interval, the operation of the instrument (screen content change, etc.) still triggers readings that make the sampling frequency uneven. The AirVisual Pro that was the source of data used in this paper is from the first generation and the measurements were taken by the default mode.

Figure 1. One of the AirVisual nodes operated by the Department of Meteorology, Eötvös Loránd University (photo by B. Atfeh).

2.2. Measurements

In December 2016 an AirVisual Pro was deployed inside a residential apartment in Budapest (19th District), Hungary to monitor $PM_{2.5}$ and CO_2 concentrations. The size of the apartment is ~50 m^2 with 4 habitants. The apartment has two rooms, and AirVisual Pro was running in one of the two rooms. Ventilation within the apartment is managed by opening the windows about 3–4 times a day during wintertime. The apartment is located in an area were the main sources of $PM_{2.5}$ is the combustion process from surrounding houses used in heating systems (traffic is typically low).

In this paper data registered during January 2017 were analyzed that was characterized by extremely high air pollution in terms of $PM_{2.5}$ and PM_{10}[1].

2.3. Data processing

Five different post-processing methods were implemented in the study to calculate daily mean $PM_{2.5}$ concentration from the raw data provided by the sensor. According to *method 1,* daily

[1] https://budapest.hu/Lapok/2017/indokolt-a-riasztasi-fokozat-es-az-ahhoz-tartozo-gepjarmuforgalom-korlatozasaval-jaro-hatosagi-intezkedes-tovabbi-fennta.aspx

averages were computed from hourly averages using all available data during the specific hours. In case of *method 2,* daily means were directly calculated from all available raw data for a given day. According to *method 3,* medians are computed from the raw data using all daily readings. *Method 4* and *method 5* can be considered as data-thinning algorithms, which are based on the sampling of raw data four times in each hour. The sampling was done pseudo-randomly in case of method 4 while data were sampled in every 15 minutes in case of method 5, which is a more deterministic way to thinning the data in comparison with method 4.

Note that the simplest method for the post-processing is associated with methods 2 and 3. Through the implementation of the latter we can mitigate the effect of outliers because median is more robust to outliers than sample mean. Those are followed by method 1 in terms of complexity that is a two-step method. In the study, methods 4 and 5 have the largest computational complexity. With the application of them we aimed to take into account each data with the same weight from the raw dataset. Their usage requires advanced data-processing tools, i.e. sampling hours in which more than 4 measurements are available while hours with 4 or less measurements remain intact. In case of method 5 the closest readings to the 1st, 16th, 31st and 46th minutes in each hour were selected.

The above-described methods were implemented in the R programming language (R Core Team, 2020), which allows us to write reusable codes to easily examine other data series in future studies. After calculation of the daily $PM_{2.5}$ averages the results are evaluated against one reference method and compared with the official data.

3. Results and discussion

3.1. Outdoor conditions

During January 2017, a high air pollution episode occurred in Budapest and in Hungary in general as well. According to *Figs. 2* and *3*, the PM_{10} and $PM_{2.5}$ concentrations were extremely high mostly during the second half of the month. New Year's Day was also characterized by high pollution most likely caused by the fireworks and petards.

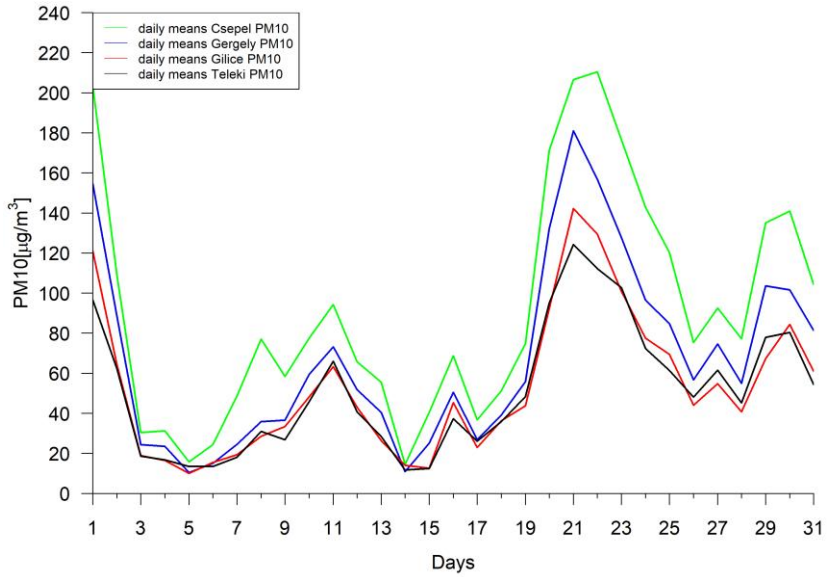

Figure 2. Daily PM_{10} concentration during January 2017 at four monitoring sites in Budapest operated by the Hungarian Air Quality Monitoring Network (source of data: www.levegominoseg.hu).

At Teleki tér and at Gilice tér the $PM_{2.5}$ and PM_{10} concentrations were very close (cf. *Figs. 2* and *3*) indicating that the primary source of the pollution was domestic heating and not traffic. As indoor air pollution is closely related to outdoor conditions due to mixing of air between indoor and the outdoor air, it is expected that indoor air quality was also extremely poor in those days in January.

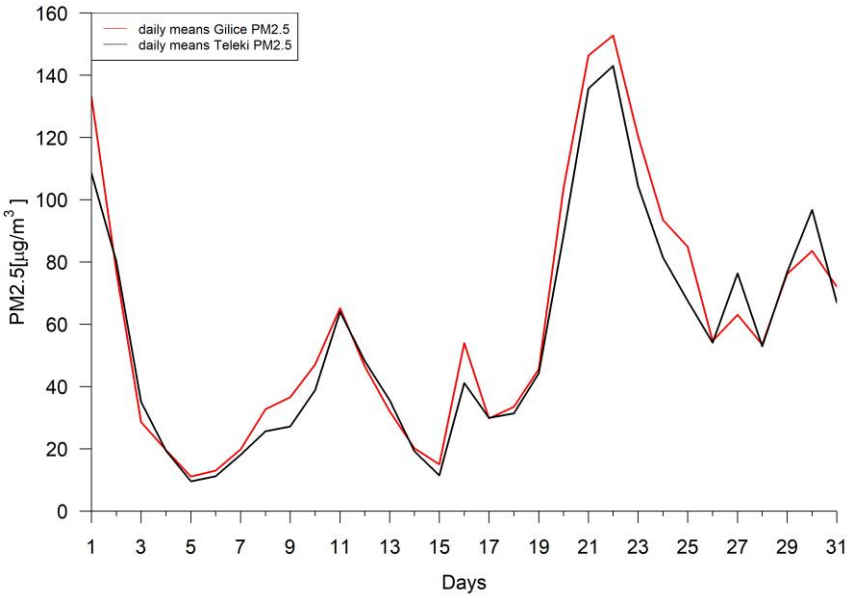

Figure 3. Daily $PM_{2.5}$ concentration during January 2017 at two monitoring sites in Budapest operated by the Hungarian Air Quality Monitoring Network (source of data: www.levegominoseg.hu).

3.2. Evaluation of indoor air quality data

In January, a total of 16,939 measurements (i.e. $PM_{2.5}$ readings) were available for a total of 735 hours. Missing hours were detected on 14th January (7 pm, 10 pm and 11 pm) and on 15th January (from 0 am to 6 am). *Fig. 4* shows the AirVisual Pro-based $PM_{2.5}$ concentration for a selected day in January 2017. As the AirVisual Pro device was operated in the so-called default mode, readings are typically available in 15-min intervals. However, during daytime the observation frequency is increasing in some occasions due to human intervention (typically when the screen is switched on or the screen content is adjusted). Note that the increase of the observation frequency co-varies with the $PM_{2.5}$ concentration in some cases which is the clear indication of interest in the actual $PM_{2.5}$ concentration during ventilation (i.e. when the windows are open).

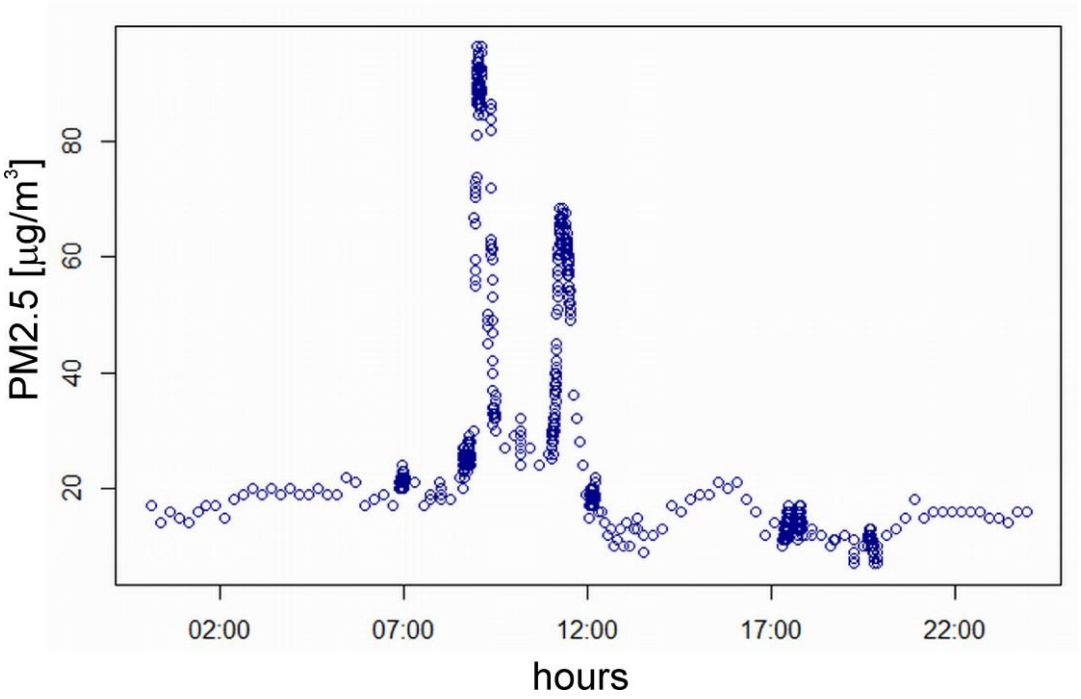

Figure 4. Indoor $PM_{2.5}$ concentration during 25 January 2017, as observed by the AirVisual Pro sensor.

Fig. 5 shows the frequency distribution of the number of readings per hour for the whole month. As it can be seen in the figure, frequency of the number of readings per hour is not uniformly distributed.

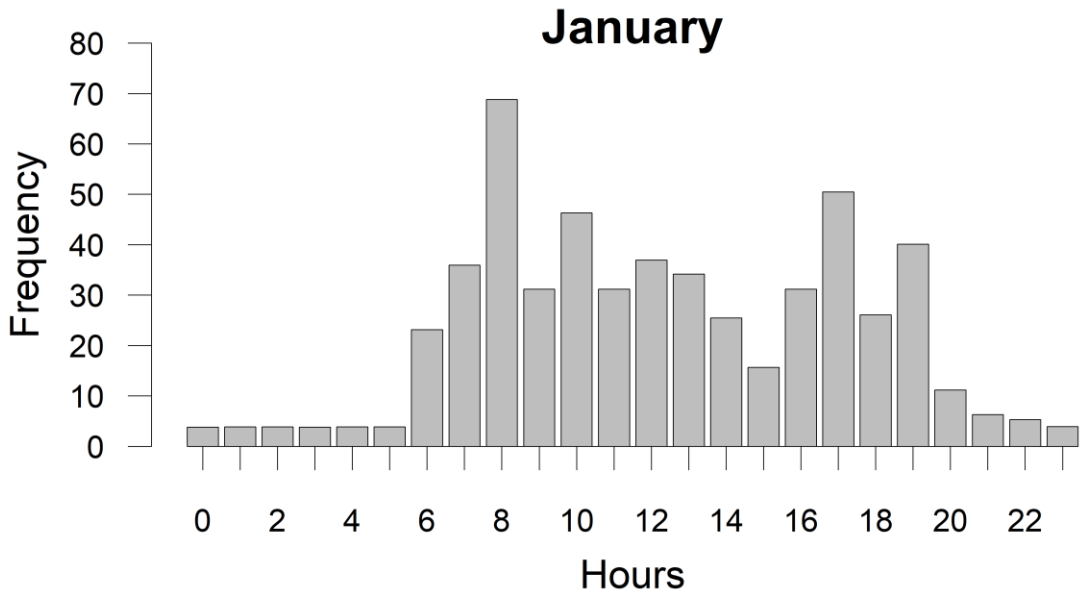

Figure 5. Frequency distribution of the number of readings per hour during January 2017.

During nighttime, when the AirVisual node is not used, the observations are made every 15 minutes (only 13 hours from a total of 735 are characterized with less than four measurements per hour, while 376 hours are available with four measurements and more than four readings are available in 346 hours). Consequently, it is expected that daily averages computed from the raw data and daily averages computed from the hourly data will not be the same. If daily averages are computed from the raw data then the averages might be distorted by hours in which larger number of readings is available. If daily averages are computed from hourly averages then each hour is taken into account with the same weight.

Fig. 6 shows the daily $PM_{2.5}$ concentration for January 2017, based on the 5 different methods (see above). Overall, the monthly course follows the outdoor conditions quite well including the high episode in 1 January (see *Fig. 3*). The period between 19 and 26 January was characterized by very high indoor $PM_{2.5}$ concentration up to around 50–60 μg m^{-3}.

It is clear from *Fig. 6* that the daily results are affected by the data processing technique as it was expected. Especially methods 2 and 3 provide markedly different results from the other 3 methods in some days (not necessarily on days that are characterized by the highest air pollution). Methods 1, 4 and 5 provide consistent results. All three methods use hourly aggregation before calculation of the daily data.

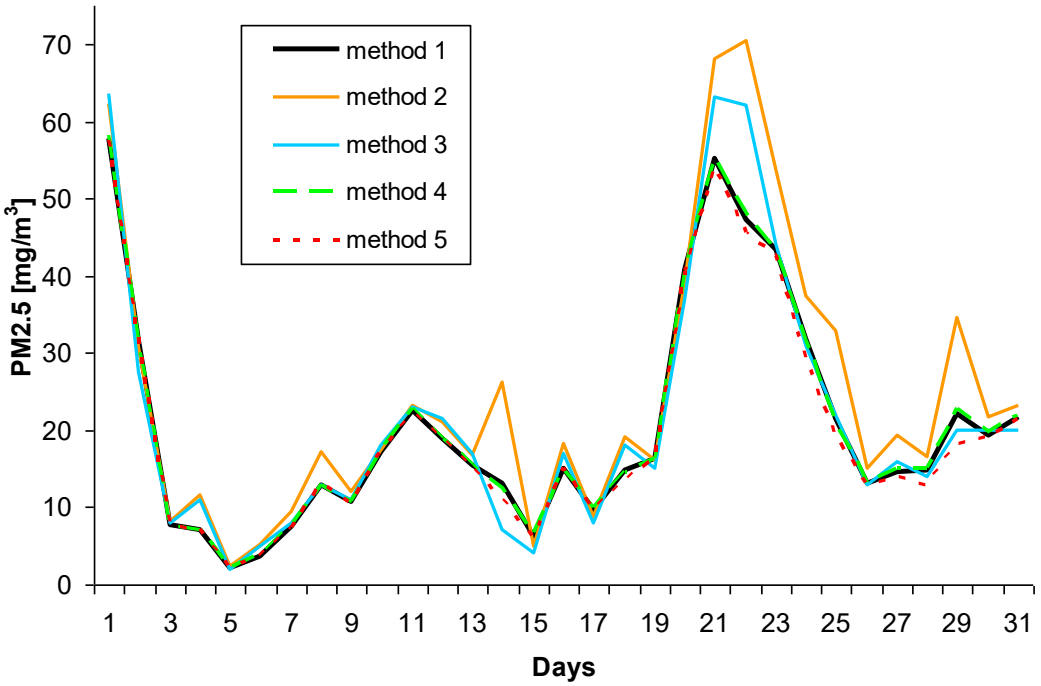

Figure 6. Daily indoor $PM_{2.5}$ concentration in the study period calculated with the different methods (see text for details).

Based on the theoretical considerations mentioned above we selected method 1 as the reference technique because it uses all available observations, and moreover it handles the possible bias caused by the high number of readings during daytime (when episodes may occur).

Table 1 provides statistical evaluation of the daily $PM_{2.5}$ data based on the five methods. Monthly mean indoor $PM_{2.5}$ concentration was very similar for methods 1, 3, 4 and 5 with negligible differences (within 1 μg m^{-3}). However, application of method 2 resulted in a large bias (4.1 μg m^{-3}). Considering the daily data the maximum difference between method 2 and method 1 was 23.2 μg m^{-3} (on 22nd January), while for method 3 and method 1 it was

14.9 μg m^{-3} (same day). Underestimation of daily $PM_{2.5}$ relative to method 1 was not typical, though for method 3 it was 6.1 μg m^{-3} in 14 January. RMSE was the largest for method 2, with somewhat smaller values for method 3. The smallest RMSE was associated with method 4 (random sampling). Explained variance relative to method 1 was quite high for all cases. The lowers explained variance (expressed by R^2) was again associated with method 2.

Table 1: Statistical evaluation of the daily $PM_{2.5}$ dataset that was derived from the raw AirVisual Pro readings using different methods. SD stands for standard deviation, RMSE is root mean square error, while R^2 is the square of the linear correlation coefficient. Bias, RMSE and R^2 values are calculated relative to method 1.

	method 1	**method 2**	**method 3**	**method 4**	**method 5**
MEAN (μg m^{-3})	20.5	24.6	21.3	20.6	19.8
SD (μg m^{-3})	14.61	17.99	16.64	14.69	14.42
BIAS (μg m^{-3})	N/A	4.1	0.7	0.1	–0.7
RMSE (μg m^{-3})	N/A	6.83	3.80	0.28	1.19
R^2	N/A	0.924	0.949	1.00	0.995

The statistical evaluation revealed that the two-step methods (methods 4 and 5) are suitable for the calculation of the daily means and provide very similar results to method 1, while the simple daily aggregation (methods 2 and 3) are associated with errors thus should be avoided. Given the simplicity of method 1 relative to methods 4 and 5 we propose to use method 1 as the 'best practice' for the post-processing of the AirVisual Pro readings. This method might be applicable to any other sensor that is characterized by non-uniform observation frequency.

4. Conclusions

LCSs provide essential information about the indoor air quality that is a major step forward in terms of the improvement of life quality and human health. According to the growing literature, AirVisual Pro is among the better LCSs that might be suitable for air pollution exposure assessments. Rigorous evaluation of the performance of the LCSs is a prerequisite for any scientifically sound assessment. In this study we demonstrated that post-processing method selection is another vital question that needs attention, especially if the observation frequency of the instruments is not constant. We propose to use a two-step data processing technique for AirVisual Pro that first consists of the calculation of hourly averages from all available readings, and then the calculation of the daily means from the hourly data. Simpler data handling (e.g. simple daily aggregation based on all raw data) might lead to inaccurate results that will affect the exposure assessment and any further conclusion regarding indoor air quality.

In our study the validity of the $PM_{2.5}$ readings was not addressed due to the lack of reference observation for the investigated time period. Evaluation of the AirVisual Pro's readings against reference monitor and adjustments (i.e. calibration) is needed in the future to provide high quality data for indoor air quality assessment.

Acknowledgements

This work was supported by the National Research, Development and Innovation Office of Hungary (grant No. 128805 and 128818) and the by the Széchenyi 2020 program, the European Regional Development Fund, and the Hungarian Government (GINOP-2.3.2-15-2016-00028). Supported by grant "Advanced research supporting the forestry and wood-processing sector´s adaptation to global change and the 4th industrial revolution", No. CZ.02.1.01/0.0/0.0/16_019/0000803 financed by OP RDE"

References

Bulot, F.M.J., Russell, H.S., Rezaei, M., Johnson, M.S., Ossont, S.J.J., Morris, A.K.R., Basford, P.J., Easton, N.H.C., Foster, G.L., Loxham, M., Cox, S.J., 2020: Laboratory Comparison of Low-Cost Particulate Matter Sensors to Measure Transient Events of Pollution. *Sensors,* 20: 2219. https://doi.org/10.3390/s20082219

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope III, C.A., Apte, J.S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., et al., 2018: Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences,* 115: 9592–9597. https://doi.org/10.1073/pnas.1803222115

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Der Boghossian, B., Cocker, D., Polidori, A., 2019: Performance evaluation of twelve low-cost $PM_{2.5}$ sensors at an ambient air monitoring site. *Atmospheric Environment,* 216: 116946. https://doi.org/10.1016/j.atmosenv.2019.116946

Joss, M.K., Eeftens, M., Gintowt, E., Kappeler, R., Künzli, N., 2017: Time to harmonize national ambient air quality standards. *International Journal of Public Health,* 62: 453–462. https://doi.org/10.1007/s00038-017-0952-y

Kistler, M., Schmidl, C., Padouvas, E., Giebl, H., Lohninger, J., Ellinger, R., Bauer, H., Puxbaum, H., 2012: Odor, gaseous and PM10 emissions from small scale combustion of wood types indigenous to Central Europe. *Atmospheric Environment*, 51: 86–93. https://doi.org/10.1016/j.atmosenv.2012.01.044

Leung, D.Y., 2015: Outdoor-indoor air pollution in urban environment: challenges and opportunity. *Frontiers in Environmental Science*, 2:, article 69. https://doi.org/10.3389/fenvs.2014.00069

Li, J., Mattewal, S.K., Patel, S., Biswas, P., 2020: Evaluation of nine low-cost-sensor-based particulate matter monitors. *Aerosol and Air Quality Research,* 20: 254–270. https://doi.org/10.4209/aaqr.2018.12.0485

Lowther, S.D., Jones, K.C., Wang, X., Whyatt, J.D., Wild, O., Booker, D., 2019: Particulate matter measurement indoors: A review of metrics, sensors, needs, and applications. *Environmental Science & Technology,* 53: 11644–11656. https://doi.org/10.1021/acs.est.9b03425

Majd, E., McCormack, M., Davis, M., Curriero, F., Berman, J., Connolly, F., Leaf, P., Rule, A., Green, T., Clemons-Erby, D., Gummerson, C., 2019: Indoor air quality in inner-city schools and its associations with building characteristics and environmental factors. *Environmental Research,* 170: 83–91. https://doi.org/10.1016/j.envres.2018.12.012

R Core Team, 2020: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Tagle, M., Rojas, F., Reyes, F., Vásquez, Y., Hallgren, F., Lindén, J., Kolev, D., Watne, Å.K., Oyola, P., 2020: Field performance of a low-cost sensor in the monitoring of particulate matter in Santiago, Chile. *Environmental Monitoring and Assessment,* 192: 71. https://doi.org/10.1007/s10661-020-8118-4

Zamora, M.L., Rice, J., Koehler, K., 2020: One-year evaluation of three low-cost $PM_{2.5}$ monitors. *Atmospheric Environment,* 235: 117615. https://doi.org/10.1016/j.atmosenv.2020.117615

Online references

AQ-SPEC, 2018: Field Evaluation IQAir AirVisual Pro (v1.1683) Sensor. http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/iqair-airvisual-pro-(fw1-1683)field-valuation.pdf?sfvrsn=8 [visited in 10 Dec, 2020]

EMEP, 2020: Center on emission inventiores and projections. https://www.ceip.at/data-viewer [visited in 10 Dec, 2020]

IQAir website, 2020a: AirVisual Pro Tech Specs. https://www.iqair.com/support/tech-specs/airvisual-pro [visited in 08 Dec, 2020]

IQAir website, 2020b: What technologies contribute to making AirVisual sensors the most accurate among low-cost monitors? https://support.iqair.com/en/articles/3562521-what-technologies-contribute-to-making-airvisual-sensors-the-most-accurate-among-low-cost-monitors [visited in 08 Dec, 2020]

IQAir website, 2020c: How often does the AirVisual Pro's sensor take measurements? What is the difference between "Continuous" and "Default" mode? https://support.iqair.com/en/articles/3029368-how-often-does-the-airvisual-pro-s-sensor-take-measurements-what-is-the-difference-between-continuous-and-default-mode [visited in 07 Dec, 2020]

WHO, 2018: World Health Organization factsheet. Household air pollution and health. https://www.who.int/en/news-room/fact-sheets/detail/household-air-pollution-and-health [visited in 01 Dec, 2020].

ORCID

Atfeh B. https://orcid.org/0000-0002-9543-3727
Kristóf E. https://orcid.org/0000-0001-9892-9552
Mészáros R. https://orcid.org/0000-0002-0550-9266
Barcza Z. https://orcid.org/0000-0002-1278-0636